Using a Human Drug Network for Generating Novel Hypotheses about Drugs

Hossein Rahmani [1a], Hendrik Blockeel[b,c], Andreas Bender[d]

[a] *Department of Knowledge Engineering, Universiteit Maastricht, PO Box 616 6200 MD, Maastricht, The Netherlands*
[b] *Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium*
[c] *Leiden Institute of Advanced Computer Science, Leiden University Niels Bohrweg 1, 2333 CA Leiden, The Netherlands*
[d] *Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

Abstract

Analyzing different drugs for various purposes is an important issue in the area of computational biology. We categorize the previous computational studies into Individual and Network approaches. While the Individual approach focuses on one specific drug without considering its relationship with other drugs, the Network approach considers also the drugs relationships. In this paper, we apply a Network approach, previously proposed for discovering the relationships among diseases, to drug data. We construct a Human Drug Network (HDN) for 200 different drugs based on functional and structural information available in the PPI network. For evaluating our proposed HDN, first, we analyzed the literature to prove that the proposed HDN is biologically meaningful. Second, we used the HDN to augment the initial prior knowledge of different drugs. As an example of prior knowledge, we considered the initial seed proteins (a set of proteins which are previously known to be drug targets) of each drug. We clustered the HDN nodes using the Markov CLustering Algorithm (MCL) and then, we augmented the seed proteins of each drug based on the cluster it belongs to. In the end, we concluded that our proposed HDN enables us to generate novel hypotheses (in terms of potential drug target proteins) and produce complementary results comparing to existing methods.

[1]Corresponding author: H.Rahmani@MaastrichtUniversity.nl

Keywords: Human Drug Network, Drug-target Proteins, Protein-Protein Interaction Networks.

1. Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [1]. Much can be learned from the analysis of such networks with respect to the metabolic and signaling processes present in an organism, and the knowledge gained can also be prospectively employed e.g., for the task of protein function prediction [2, 3, 4], identification of functional modules [5], interaction prediction [6, 7], identification of disease candidate genes [8, 9, 10, 11, 12, 13, 14, 15, 16, 17] and drug targets [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], according to an analysis of the resulting network [32].

It is estimated that a typical drug discovery cycle, from target identification to clinical use, can take 14 years [33] with cost of 800 million US dollars [34]. Recently, computational methods have been widely applied to facilitate the design and discovery of new drugs [35]. Identification of new drug target proteins is the first and still very important step in the drug discovery cycle [36]. Although computational methods can not substitute the real screening targets of a novel drug, they succeed to lower the costs and time in all steps of the drug discovery cycle. They investigate huge amounts of chemical compounds which may have not been synthesized before [37] and then, they generate novel hypotheses about newly-identified drug target proteins and accordingly, reduce the initial number of compounds for the experimental analysis.

We categorize the previous computational studies on drug-target identification into Individual and Network approaches. The Individual approach focuses on one specific drug without considering its relationship with other drugs. Methods implementing this approach usually transform different properties of target proteins into numeric features and then, they apply computational methods to detect potential drug-target interactions [18, 19, 32, 38, 39, 40, 41, 42, 43]. The main difference among the Individual methods lies in the type of properties they consider for the prediction. Complementary to Individual approach, the Network approach first discovers the hidden relationship among different drugs, and then it uses this information to predict novel drug target proteins by sharing information across highly-related

drugs [29, 44, 45, 46, 47]. The main assumption behind the Network approach is that similar drugs tend to target similar proteins [48, 49]. The method proposed by Yildirim et al., [29], which has a prominent role among the network based approaches, connects two drugs in the network if they share at least one target protein. However their approach is not capable of discovering relationships among drugs with no common target proteins. In this paper, we apply the existing Network approach for discovering the relationships among diseases [50] on drug data. In the resulting network, two drugs even with no common target proteins may be connected to each other if they are found to be similar with respect to contextual information extracted from PPI networks. Once the Human Drug Network (HDN) is constructed, the information in the HDN can be used to generate novel hypotheses about drug targets, identifying genes that may be involved in some drugs but would not be detected using previous approaches.

Section 2 discusses the previous methods on drug target prediction. We describe the network approach in detail in Section 3. In Section 4, the method is applied to a concrete PPI network; using functional and structural information from the network, a drug network is constructed for 200 drugs. This network is next evaluated in terms of interpretability and knowledge augmentation. Section 5 concludes.

2. Background

The identification of relations between drugs and their target proteins is a key area in drug discovery cycle [36]. Complementary to high-throughput screening methods, computational approaches are used extensively to reduce costly late-stage drug failures and accelerate successful development of new drugs.

For the task of drug target prediction, existing computational approaches could either focus on each drug individually and neglect the relationships among drugs [18, 19, 32, 38, 39, 40, 41, 42, 43] or they could consider the informative relationships among the highly-related drugs in the prediction process [29, 44, 45, 46, 47]. The former and the latter approaches are called Individual and Network approaches, respectively.

The Individual approach predicts drug targets for a individual drug according to the drug's informative features. Zoraghi et al., [18] argue that the architecture of bacterial or hostpathogen protein interactomes can provide invaluable insights for the identification of novel antibacterial drug targets. De

Las Rivas et al., [19] find novel drug targets according to the location of the proteins in the interactome network. Maayan et al. [32] constructed a bipartite network connecting drug targets and drugs and subsequently analyzed the targets in the context of a global protein-protein interaction network. Keiser et al., [38] compare 3,665 US Food and Drug Administration (FDA)-approved and investigational drugs against hundreds of targets, defining each target by its ligands and then they predict thousands of unanticipated associations according to chemical similarities between drugs and ligand sets. Campillos et al., [39] predict new targets for established drugs looking for side-effects shared between two molecules. Kolarik et al., [40] developed an approach for the identification of new terms used in unstructured text that provide information about drug properties. Hert et al., [41] build target relationships based on the structural and biological similarity of their ligands. Hwang et al., [42] propose an approach for identifying 'bridging nodes' in the biological network as potential drug targets. Bleakley et al., [43] introduce a Bipartite Local Models (BLM) to first predict target proteins of a given drug, then to predict drugs targeting a given protein.

The Network approach, mostly, assumes that similar drugs tend to target similar proteins [48, 49]. Yildirim et al., [29] connect two drugs if they share a common target. They show that most new drugs interact with previously targeted cellular components and there are relatively few drugs entering the market with novel targets. There is a family of methods which use chemical structural information to build drug-drug relationships [44, 45, 46, 47]. They first convert the chemical structure into numerical vectors and then define the drug-drug relationship according to vector similarity.

3. Materials and Methods

3.1. Terminology and Symbols

We consider a PPI network as an undirected annotated graph $(P, E, \lambda_F, \lambda_D)$ where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ_F and λ_D are so-called annotation functions; for each protein p, λ_F and λ_D denote additional information we have about p. $\lambda_F(p)$ lists all the GO functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. $\lambda_D(p)$ lists all the drugs that target protein p; we call it the drug list of p and denote it $drugList(p)$. We also define seed proteins $SP(d_i)$ as the set of proteins targeted by drug d_i ($d_i \in drugList(p) \Leftrightarrow p \in SP(d_i)$).

3.2. Human Drug Network

We now describe our method for building a Human Drug Network (HDN) from a PPI network. Essentially the same method was used before to construct human disease networks, with good results [50]. We define a Human Drug Network (HDN) as a directed graph $HDN(D, R)$ where D is a set of drugs and $R \subseteq D \times D$ is a set of directed relationships between these drugs. We build our proposed HDN as follows.

For each drug d_i, we learn a model that can predict, for any protein p, how likely p is targeted by this drug. Next, we use this model to make predictions for all the seed proteins of a drug d_j. The higher these seed proteins score, on average, the stronger the link between d_i and d_j is considered to be.

Concretely, the model for a drug d_i is learned and used as follows:

1. let $testSet$ contain the seed proteins of all drugs except d_i.
2. let $trainSet$ contain all proteins not in $testSet$
3. We learn a predictive model M from $trainSet$, using the seed proteins of d_i as positive examples and all other proteins as negative examples. We then use M to predict for each protein in $testSet$ how likely it is targeted by d_i (higher values meaning more likely). For randomized learners, we repeat this 10 times (otherwise just 1 time) and calculate for each $p \in testSet$ the average, denoted $APV(p)$.
4. For each drug $d_j \in D(j \neq i)$, we add a directed edge $d_i \rightarrow d_j$ to the HDN with a weight

$$weight(d_i \rightarrow d_j) = \frac{\sum_{p \in SP(d_j)} APV(p)}{|SP(d_j)|} \tag{1}$$

with $|SP(d_j)|$ the number of seed proteins of d_j.

This procedure is repeated for all drugs. The resulting HDN is a directed, fully connected network in which each node is a drug and each weighted edge shows a relationship between two drugs. A high weight for $d_i \rightarrow d_j$ expresses that proteins targeted by d_j are, on average, likely to be targeted also by d_i, according to the model built for d_i.

In order to focus on the most important relationships in the HDN, we prune the network by keeping only the highest-ranked edges. Section 4.2 discusses the pruning procedure of the original network in details.

There are many ways in which the predictive model M can be learned from the PPI network (step 3). Based on the study on the same PPI dataset

[50], we choose a hybrid prediction method which considers both Structural and Functional information in the PPI network. We discuss this method in the next section.

3.3. Hybrid Prediction Method

Rahmani et al., [50] analyzed different functional and structural prediction methods and they observed that a hybrid method that considers both functional and structural information in the PPI network worked best for building the Human Disease Network. In this paper, we use the same prediction method for building Human Drug Network. The following section describes the proposed structural, functional and the hybrid prediction methods in details.

3.3.1. Structural Information (ST-RW)

Berger et al. [51] propose a random walk based approach to predict disease-related proteins in PPI networks. They assume that disease-related proteins fall closer, with respect to average number of steps a random walker takes to walk from a specified protein to another one, to the seed proteins than they do on average to the rest of the proteins in the PPI network. They calculate the score of each protein p_j in the PPI network based on Formula 2 and then, select high-scoring proteins as disease-related proteins.

$$score_s(p_j) = \frac{\frac{\sum_{i \in C'} T_{ij}}{|C'|} - \frac{\sum_{i \in C} T_{ij}}{|C|}}{\frac{\sum_i T_{ij}}{|C|+|C'|}} \tag{2}$$

In Formula 2, T_{ij} is the average number of steps a random walker takes to walk from a specified node i to another specified node j, C is the set of seed proteins and C' is the set of all other proteins in the network. In the rest of this paper, we refer to this method as *ST-RW*.

3.3.2. Functional Information (Func-Indiv)

This method uses the functional annotations of proteins. For each function, it determines how strongly the function correlates with drug-target relatedness, using the standard χ^2 statistic as proposed by Liu et al. [52]:

$$\chi^2(f_i) = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{3}$$

where $a = |D \cap P_i|$, $b = |D \cap \bar{P}_i|$, $c = |\bar{D} \cap P_i|$, and $d = |\bar{D} \cap \bar{P}_i|$, with P_i and $\bar{P}_i$ the set of proteins labeled and not labeled with function f_i, and D and $\bar{D}$ the set of drug-target and not drug-target proteins, respectively.

Next, it describes all proteins using the highest-scoring functions as features, and then a standard machine learning system can be used to learn a model that from these informative features predicts the likelihood of involvement. We here use Naive Bayes [53]. This method estimates the conditional probability distribution for the variable to be predicted, given the feature values, and up to a constant factor, as follows:

$$p(C|F_1, \ldots, F_n) \sim p(C)p(F_1|C)p(F_2|C)\cdots p(F_n|C) \tag{4}$$

This estimation of the conditional probability distribution relies on conditional independence of the features given the target. This assumption is usually violated, but the method is quite robust to violations of the assumption [54], and works well in practice. Furthermore, several researchers found, in a similar context, that the features are more important than the actual machine learning method used [55]. In the rest of this paper, we refer to this method as *Func-Indiv*.

3.3.3. Integrating Functional and Structural Information

Structural-based and functional-based methods can be combined into a hybrid method. The hybrid method is calculated as follows:

$$score_h(p) = norm(score_s(p)) + norm(score_f(p)) \tag{5}$$

In Formula 5, $score_s(p)$ and $score_f(p)$ represent the drug-relatedness score of p using a *Structural* (*ST-RW*) and a *Functional* (*Func-Indiv*) method, respectively. In order to avoid a bias toward either of these categories, we normalize the drug-relatedness scores using

$$norm(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{6}$$

where $min(x)$ and $max(x)$ return the minimum and maximum values taken over all values of x, respectively. In the rest of this paper, we refer to this method as *RW-Indiv*.

4. Empirical Results and Discussion

In this section, we discuss several aspects of the proposed method in more detail, and investigate them experimentally.

The dataset used for the experiment is described in Section 4.1. In Section 4.2, we show the HDN constructed by our method, and provide a biological interpretation; this is meant as an evaluation of how informative and interpretable the network is. Finally, a more objective type of evaluation is: How useful is the HDN from the point of view of predicting drugs targets? Does using the network yield better predictions? In section 4.3, we augment the initial seed proteins of different drugs by sharing information across highly-related drugs. Then, we evaluate the use of augmented seed proteins for predicting drug target proteins in two different ways.

4.1. Dataset

We applied our method for building the HDN to the PPI network used by Milenkovic et al. [56]. This dataset is the union of three human PPI datasets: HPRD [57], BIOGRID [58] and the dataset used by Radivojac et al. [59] and contains $47,303$ physical interactions among $10,282$ proteins. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. Milenkovic et al. [56] provide details on the construction of the integrated network. The GO functions of proteins are extracted from [60]. Table 1 shows some basic statistical information about our annotated dataset.

We analyzed 200 drugs from the DrugBank [61] database. Table 2 shows the 20 drugs with the highest number of seed proteins. The average number of seed proteins for all 200 drugs is 11.51. The list of 200 drugs with their seed proteins is listed in `http://www.liacs.nl/~hrahmani/HDN/drugtargets.html`

4.2. Novel Human Drug Network

The hybrid prediction method discussed in Section 3.3 rely on a feature selection step, for which a number of features needs to be decided. Based on earlier work on the same dataset [62], we consistently choose 100 functional and 10 structural features. Having made this choice, we build our proposed

HDN for 200 different drugs. There are 39800 (200×199) possible edges in the original HDN. Each edge $d_i \rightarrow d_j$ shows the average rank of seed proteins of d_j among all the proteins in the Test set using the seed proteins of d_i as positive examples in the Train set. Train and Test sets are both described in Section 3.2. The lower average rank indicates stronger relationship between d_i and d_j, in comparison with other drugs. To select the most informative relationships among drugs, first, we sort the edges according to their score (average rank of seed proteins), ascendingly. The result is shown in Figure 1. In this figure, the X axis shows the 39800 edges in the HDN and the Y axis shows, for each edge $d_i \rightarrow d_j$, the average rank of the seed proteins of d_j (the smaller, the better). Second, we determine the candidate cutoff points by discovering two turning points in the curve, roughly at 3% and 86% of all edges. Finally, instead of analyzing the whole HDN, we focus on the pruned HDN containing only the 1328 (3% of the original HDN) highest-ranked edges. This turns the fully-connected graph into a more informative, visualizable graph.

Figure 2 shows the pruned HDN using cytoscape [63]. Then, to discover highly-related drugs in the pruned network we cluster the pruned HDN using the Markov Clustering (MCL) [64]. We should emphasize that in our pruned Human Drug Network, each edge shows an informative relationship between two drugs and accordingly, the clustering method should focus on network connections as main indicators of related drugs. The key intuition behind Markov Clustering (MCL) [64] is that a "random walk that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited". MCL has been applied in a number of different domains, mostly in bioinformatics [65, 66, 67, 68]. Figure 3 and Table 3 represent 13 clusters extracted from the pruned network using MCL clustering in graph and tabular formats, respectively.

We will now briefly discuss the biological significance of the observed findings in Figure 2. Clearly visible are, from left to right, firstly a group of both strong painkillers and hypnotics (e.g. tramadol) and analogues of ephedrine. This group also contains many compounds with high abuse potential (amphetamines, cocaine, tramadol) which often work on the opinion receptors, as well as the serotonin transporter system. The next large group, containing e.g. methamphetaime, moves away from activity on the opioid receptors, and generally to more polypharmacology again neurotransmitter transporters. The ethchlorvynol group that comes next is a classical hypnotics group, with barbital action generally being allosteric modulation of

GABA-A receptors. It is interesting - and correct - to note that the benzodiazepine group is located next to it (containing e.g. diazepam, better known under its trade name Valium) and hence strongly linked via the biological targets modulated; however, separation is not complete (e.g. Aprobarbital is here contained within the benzodiazepine group). The following - disjoint - group is concerned with matrix remodeling, involving proteins such as urokinase (as well as Vitamin C, where lack of it is known to involve tissue disorder, scorbut). Further group contain kinase inhibitors used to treat cancer (e.g. dasatinib), as well as an antibody group (containing e.g. etanercept). Hence, overall, we can conclude that both the groups themselves, as well as their relationships to groups involving similar biological processe, are overall biologically rather meaningful.

In addition to known bioactivities for drugs, after applying the MCL algorithm to clustering drugs, a different - and possibly cleaner - picture of drug bioactivity emerges. This picture could provide insights for exploring the new aspects of drug bioactivities. As shown in Figure 3, both barbiturates and diazepams now cluster together (Cluster 1 in Table 3), which is understandable given that both act to a good extend on the GABA-A receptor. Also drugs working on other GPCRs in the brain (a well as neurotransmitter transporter, such as the one of dopamine and serotonin) show a cleaner cluster, as indicated as the next cluster (Cluster 2 in Table 3). While the kinase antibodies (Cluster 3 in Table 3) cluster also here, the same is interestingly not true for small-molecule kinase inhibitors: Here the sunitinib cluster (Cluster 6 in Table 3) contains both small molecule inhibitors of kinases as well as Adenosine Triphosphate (which is required for kinase activity), and also the group of tramadol and cocaine which could be seen in Figure 2 is now not present anymore.

4.3. Prediction from Augmented Seed Proteins

We now consider the task of drug target hypotheses generation, using a predictive model learned from data. The quality of such a predictive model obviously depends on the prior knowledge incorporated in the data. While earlier models focused on knowledge about one drug, we hypothesize that augmenting the prior knowledge about one drug with knowledge obtained from studying related drugs may yield better predictive models.

To evaluate this hypothesis, we have used our proposed HDN for augmenting the seed proteins of different drugs as follows.

First, we clustered the pruned HDN into n clusters $C_1 \ldots C_n$ using MCL clustering. The reason we chose MCL clustering and the biological meaningfulness of cluster results are all discussed in Section 4.2. Second, we augmented the seed proteins of each drug by adding the seed proteins of all the drugs in the same cluster C_j; that is, $d_i \in C_j \Rightarrow Aug(d_i) = \cup_{d_k \in C_j} SP(d_k)$, with $Aug(d_i)$ the augmented list of seed proteins of drug d_i.

Next, we have evaluated the effect of using the augmented list to learn a model, instead of the original list, along two dimensions: predictive accuracy, as well as biological meaningfulness.

First, we compare the predictions of a model learned using augmented seed proteins (Network approach) with those of a model learned using only the initial seed proteins (Individual approach). In both cases, the same learning method was used (RW-Indiv).

Figure 4 compares the Network approach with the Individual approach for 13 drug clusters augmented by our HDN, in terms of the following leave-one-out cross-validation procedure, proposed earlier by De Bie et al. [69].

For each cluster c_j

1. For each drug $d_i \in c_j$:
   (a) Randomly select 99 proteins that are not seed proteins for d_i
   (b) For each seed protein p of d_i:
      i. let $testSet$ contain p and the 99 proteins
      ii. let $trainSet$ contain all other proteins
      iii. Learn a model M from $trainSet$, apply it to $testSet$, and check how high p ranks.
   (c) Repeat steps 1a to 1b ten times and calculate the average rank $AR(p)$ of each seed protein p. The overall rank of d_i, for a given method, is then defined as

$$overallRank(d_i) = \frac{\sum_{p \in SP(d_i)} AR(p)}{|SP(d_i)|} \tag{7}$$

2. Finally, calculate the average rank of cluster c_j using Formula 8.

$$AverageRank(c_j) = \frac{\sum_{d_i \in c_j} overallRank(d_i)}{| \cup_{d_i \in c_j} |} \tag{8}$$

where $| \cup_{d_i \in c_j} |$ returns the number of drugs belong to cluster c_j.

The comparison result is shown in Figure 4. The Network approach outperforms the Individual approach in clusters 2, 6, 10 and 12 with 39, 7, 4

and 2 drug members. In total, the Network approach predicts more accurate results in 26% (52 drugs) of the 200 considered drugs. This suggests that our proposed Network approach should not be seen as replacing the Individual approach, but as complementary to it.

Beside the numerical evaluation mentioned above, we also want to evaluate to what extent the proposed network approach predicts biologically meaningful results. To this aim, we ask the domain experts to biologically interpret the proteins predicted to be targeted by drugs even when they were not annotated as such in the training data. Our approach consists of the following steps: First, we build a new training set containing the augmented list of seed proteins (positive set) in addition to 100 randomly selected proteins (negative set). Even though we are not sure that all of these random proteins are negative, it is very likely that the majority of them are negative. The remaining positive cases constitute noise in the training set. Second, we build a test set containing all the remaining proteins in the network. Third, we use *RW-Indiv* for predicting new proteins targeted by drugs. Table 4 shows 10 highest-ranked proteins predicted for each cluster shown in table 3.

In table 4, clear differences as to the bioactivity profiles of the different clusters can be observed. Cluster 1 is a classical GPCR cluster, with activity involving both serotonin and nicotinic acetylcholine receptors. Cluster 2 is related, however containing the adenosine receptors as GPCR receptors, as well as the glutamate and alpha subtypes of the nicotinic acetylcholine receptors. With cluster 3 we are moving into the area of peptide receptors, such as DPP4 and CCL5. Targets in cluster 4 are related to metabolism (e.g. APOE), as well as blood coagulation (THBS1, FGB) which is consistent with the drugs that belong to this cluster. Cluster 5 is a very different cluster, namely use that contains a large number of six different isoforms of the metabolizing enzyme P450, but also other related proteins (e.g. pyruvate carboxylase). In cluster 6 we move to more cancer-related targets, involving both oncogenes (e.g. AKT2), as well as growth factors (TGFBR2). While cluster 7 is difficult to get a homogeneous picture of, cluster 8 is very much dominated by the dopamine receptor subtypes (D1-D4), which are classical targets of antidepressants or, more general, CNS-active drugs. Cluster 9 also contains some of the dopamine receptors (namely D1-D3) however it also has as the highest-ranked target the serotonin 1B receptor (HTR1B) on the list. Cluster 10 is significantly different in nature - it contains solely ribosomal proteins, so those involved in protein synthesis. While cluster 11 contains both potassium and calcium channels (many of which are involved

in the action potential of the heart), cluster 12 contains also ion channels, but mostly different ones from the ones in the previous cluster. The same can be said about Cluster 13, which now involves also a variety of sodium channels in addition to the above.

The biological meaningfulness of the result confirms the possibility of using the proposed Human Drug Network for generating other types of novel hypotheses for drugs. For example, we could investigate whether the co-clustering drugs in the Human Drug Network share similar side-effects or not.

5. Conclusions and Future Work

The previous studies on analyzing different drugs can be categorized into the Individual and Network approaches. While the Individual approach focuses on one specific drug without considering its relationship with other drugs, the Network approach considers also the drugs relationships. In this paper, we examine a previous Network approach for discovering the relationships among diseases [50] on drug data and we showed that our method is capable of generating novel hypotheses (in terms of complementary drug target proteins) by considering the informative relationships among drugs. We built an HDN for 200 different drugs. The discovered relationships among drugs were biologically discussed and validated by domain expert. Then, we clustered the HDN nodes using Markov CLustering Algorithm (MCL) and we augmented the seed proteins of drugs based on the cluster they belong to. Finally, we compared the predictions of models learned using the augmented list with those of models learned using the original list. We observed that our proposed method outperforms the Individual approach in 26% (52 drugs) of the 200 considered drugs.

As future work, we could improve and extend the proposed method in several directions. In the first direction, we could apply more extensive validation to the result of our proposed approach. We have already discussed and validated our results in sections 4.2 and 4.3 by domain expert, however, biological/experimental validation of the findings using methods such as PR and RT-PCR is still challenging and needs separate studies. Additionally, although we believe strongly that the results of this paper reduce the search space, generate novel hypotheses and bring new insights for the biologists and clinical researchers, these results should not be assumed as effective and employed in action by pharmaceutical companies unless they are validated experimentally in the laboratories. In the second direction, in addition to functional and structural features, we could consider other biological features in the system. In the third direction, we could investigate the complementarity nature of our proposed network approach in more depth. For instance, we could compare the proteins predicted by the Network approach (as listed in Table 4) to those predicted by the individual approach and then discuss the targets that were discovered by the network approach but missed by the individual approach. In the fourth direction, we could use the proposed Human Drug Network for generating other types of novel hypotheses. For example, co-clustering in the Human Drug Network may suggest some similarities in

pharmacodynamics of the drugs. It may generate novel hypothesis on a new drug indication. In addition, co-clustering may suggest some similarities in pharmacologic profiles of drugs. These include pharmacokinetic properties, drug interaction as well as side effect profiles.

References

[1] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E. E. Wanker, A human protein-protein interaction network: a resource for annotating the proteome., Cell 122 (2005) 957–68.

[2] T. Milenkovic, N. Przulj, Uncovering biological network function via graphlet degree signatures, Cancer informatics 6 (2008) 257–73.

[3] H. Rahmani, H. Blockeel, A. Bender, Predicting the functions of proteins in protein-protein interaction networks from global information, in: JMLR: Workshop and Conference Proceedings,, 2010, pp. 82–97.

[4] H. N. Chua, W.-K. Sung, L. Wong, Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions, Bioinformatics 22 (2006) 1623–30.

[5] Z. Lubovac, J. Gamalielsson, B. Olsson, Combining functional and topological properties to identify core modules in protein interaction networks, Proteins 64 (2006) 948–59.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, M. Gerstein, A bayesian networks approach for predicting protein-protein interactions from genomic data, Science 302 (2003) 449–53.

[7] S. Wong, L. Zhang, A. Tong, Z. Li, D. Goldberg, O. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, Combining biological networks to predict genetic interactions, Proc Natl Acad Sci USA 101 (2004) 15682–7.

[8] S. Erten, G. Bebek, R. Ewing, M. Koyutrk, Dada: Degree-aware algorithms for network-based disease gene prioritization., BioData Min 4 (2011).

[9] A. Schlicker, T. Lengauer, M. Albrecht, Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, Bioinformatics 26 (2010) i561–7.

[10] Z. Dezso, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, A. Bugrim, Identifying disease-specific genes based on their topological significance in protein networks, BMC Systems Biology 3 (2009) 36+.

[11] S. Köhler, S. Bauer, D. Horn, P. Robinson, Walking the interactome for prioritization of candidate disease genes., American journal of human genetics 82 (2008) 949–58.

[12] L. Sam, Y. Liu, J. Li, C. Friedman, Y. Lussier, Discovery of protein interaction networks shared by diseases, Pac Symp Biocomput (2007) 76–87.

[13] H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, A. Droege, K. Lindenberg, M. Knoblich, C. Haenig, A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington's disease, Mol Cell 15 (2004) 853–65.

[14] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, Bioinformatics 22 (2006) 2800–5.

[15] M. Oti, B. Snel, M. Huynen, H. Brunner, Predicting disease genes using protein-protein interactions, J Med Genet 43 (2006) 691–8.

[16] X. Wu, S. Li., Cancer Gene Prediction Using a Network Approach. Chapter 11 Mathematical and Computational Biology, Cancer Systems Biology (Ed. Edwin Wang). Series: Chapman and Hall/CRC, 2010.

[17] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, Bioinformatics 22 (2006) 2800–5.

[18] R. Zoraghi, N. E. Reiner, Protein interaction networks as starting points to identify novel antimicrobial drug targets, Current Opinion in Microbiology (2013) –.

[19] J. Las Rivas, C. Prieto, Protein interactions: Mapping interactome networks to support drug target discovery and selection, in: R. S. Larson (Ed.), Bioinformatics and Drug Discovery, volume 910 of *Methods in Molecular Biology*, Humana Press, 2012, pp. 279–96. URL: `http://dx.doi.org/10.1007/978-1-61779-965-5_12`. doi:`10.1007/978-1-61779-965-5_12`.

[20] J. Huang, C. Niu, C. D. Green, L. Yang, H. Mei, J.-D. J. Han, Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network, PLoS Comput Biol 9 (2013) e1002998.

[21] M. Zhu, L. Gao, X. Li, Z. Liu, Identifying drug-target proteins based on network features, Science in China Series C: Life Sciences 52 (2009) 398–404.

[22] H. Ruffner, A. Bauer, T. Bouwmeester, Human protein-protein interaction networks and the value for drug discovery, Drug Discov Today 12 (2007) 709–16.

[23] J. Mestres, E. Gregori-Puigjane, S. Valverde, R. V. Sole, The topology of drug-target interaction networks: implicit dependence on drug properties and target families, Mol. BioSyst. 5 (2009) 1051–7.

[24] I. Vogt, J. Mestres, Drug-target networks, Molecular Informatics 29 (2010) 10–4.

[25] M. Zhu, L. Gao, X. Li, Z. Liu, C. Xu, Y. Yan, E. Walker, W. Jiang, B. Su, X. Chen, H. Lin, The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network., Journal of drug targeting (2009).

[26] F. Hormozdiari, R. Salari, V. Bafna, S. C. Sahinalp, Protein-protein interaction network evaluation for identifying potential drug targets., Journal of Computational Biology 17 (2010) 669–84.

[27] S. K. Kushwaha, M. Shakya, Protein interaction network analysisapproach for potential drug target identification in mycobacterium tuberculosis, Journal of Theoretical Biology 262 (2010) 284 –94.

[28] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. Gibson, J. Lewis, L. Serrano, R. Russell, Systematic discovery of new recognition peptides mediating protein interaction networks, PLoS Biol 3 (2005) e405.

[29] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, M. Vidal, Drug-target network, Nat Biotech 25 (2007) 1119–26.

[30] A. Pujol, R. Mosca, J. Farrés, P. Aloy, Unveiling the role of network and systems biology in drug discovery., Trends in pharmacological sciences 31 (2010) 115–23.

[31] E. Chautard, N. Thierry-Mieg, S. Ricard-Blum, Interaction networks: from protein functions to drug discovery. A review., Pathologie-biologie 57 (2009) 324–33.

[32] A. Ma'ayan, S. L. Jenkins, J. Goldfarb, R. Iyengar, Network analysis of fda approved drugs and their targets., The Mount Sinai journal of medicine, New York 74 (2007) 27–32.

[33] S. Myers, A. Baker, Drug discovery–an operating model for a new era., Nature biotechnology 19 (2001) 727–30.

[34] J. A. DiMasi, R. W. Hansen, H. G. Grabowski, The price of innovation: new estimates of drug development costs, Journal of Health Economics 22 (2003) 151 –85.

[35] C. M. Song, S. J. Lim, J. C. Tong, Recent advances in computer-aided drug design, Briefings in Bioinformatics 10 (2009) 579–91.

[36] M. A. Lindsay, Target discovery, Nature Reviews Drug Discovery 2 (2003) 831–8.

[37] L. Weber, Current status of virtual combinatorial library design, QSAR and Combinatorial Science 24 (2005) 809–23.

[38] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet, B. L. Roth, Predicting new molecular targets for known drugs., Nature 462 (2009) 175–81.

[39] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, P. Bork, Drug Target Identification Using Side-Effect Similarity, Science 321 (2008) 263–6.

[40] C. Kolrik, M. Hofmann-Apitius, M. Zimmermann, J. Fluck, Identification of new drug classification terms in textual resources., in: ISMB/ECCB (Supplement of Bioinformatics), 2007, pp. 264–72. URL: http://dblp.uni-trier.de/db/conf/ismb/ismb2007.html#KolarikHZF07.

[41] J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea, B. K. Shoichet, Quantifying the Relationships among Drug Classes, J. Chem. Inf. Model. 48 (2008) 755–65.

[42] W.-C. Hwang, A. Zhang, M. Ramanathan, Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery., Clin Pharmacol Ther 84 (2008) 563–72.

[43] K. Bleakley, Y. Yamanishi, Supervised prediction of drug-target interactions using bipartite local models., Bioinformatics (Oxford, England) 25 (2009) 2397–403.

[44] K. Park, D. Kim, Drug-drug relationship based on target information: application to drug target identification, BMC Systems Biology 5 (2011) S12.

[45] R. D. Brown, Y. C. Martin, The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding, Journal of Chemical Information and Computer Sciences 37 (1997) 1–9.

[46] T. R. Hagadone, Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases, Journal of Chemical Information and Modeling 32 (1992) 515–21.

[47] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, Chemical Similarity Using Physiochemical Property Descriptors, J. Chem. Inf. Model. 36 (1996) 118–27.

[48] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, Similarity metrics for ligands reflecting the similarity of the target proteins., Journal of Chemical Information and Computer Sciences 43 (2003) 391–405.

[49] T. Klabunde, Chemogenomic approaches to drug discovery: similar receptors bind similar ligands, British Journal of Pharmacology 152 (2007) 5–7.

[50] H. Rahmani, H. Blockeel, A. Bender, Using a human disease network for augmenting prior knowledge about diseases, In press, Intelligent Data Analysis 19(4) (2014).

[51] S. Berger, A. Ma'ayan, R. Iyengar, Systems Pharmacology of Arrhythmias, Sci. Signal. 3 (2010) ra30+.

[52] H. Liu, R. Setiono, Chi2: Feature selection and discretization of numeric attributes, in: In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995, pp. 388–91.

[53] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification (2nd Edition), Wiley-Interscience, 2000.

[54] P. Domingos, M. Pazzani, On the optimality of the simple bayesian classifier under zero-one loss, Mach. Learn. 29 (1997) 103–30.

[55] L. Li, K. Zhang, J. Lee, S. Cordes, D. Davis, Z. Tang, Discovering cancer genes by integrating network and functional properties, BMC Medical Genomics 2 (2009) 61.

[56] T. Milenkovic, V. Memisevic, A. K. Ganesan, N. Przulj, Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data., Journal of the Royal Society, Interface / the Royal Society 7 (2010) 423–37.

[57] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. Shivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, A. Pandey, Human protein reference database as a discovery resource for proteomics., Nucleic Acids Res 32 (2004).

[58] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, Biogrid: a general repository for interaction datasets, Nucleic Acids Research 34 (2006) 535–9.

[59] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, S. D. Mooney, An integrated approach to inferring gene-disease associations in humans., Proteins 72 (2008) 1030–7.

[60] Go annotation, ???? URL: `ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz`.

[61] D. S. Wishart, C. Knox, A. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, Drugbank: a knowledgebase for drugs, drug actions and drug targets., Nucleic Acids Research 36 (2008) 901–6.

[62] H. Rahmani, H. Blockeel, A. Bender, Predicting genes involved in human cancer using network contextual information, J. Integrative Bioinformatics 9 (2012).

[63] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, G. D. Bader, Integration of biological networks and gene expression data using Cytoscape, Nature Protocols 2 (2007) 2366–82.

[64] S. van Dongen, Graph Clustering by Flow Simulation, Ph.D. thesis, University of Utrecht, 2000.

[65] A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Research 30 (2002) 1575–84.

[66] L. Li, C. J. Stoeckert, D. S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes., Genome Res 13 (2003) 2178–89.

[67] R. Sharan, I. Ulitsky, R. Shamir, Network-based prediction of protein function, Molecular Systems Biology 3 (2007).

[68] S. Brohee, J. van Helden, Evaluation of clustering algorithms for protein-protein interaction networks, BMC Bioinformatics 7 (2006) 488+.

[69] T. De Bie, L. Tranchevent, L. M. M. van Oeffelen, Y. Moreau, Kernel-based data fusion for gene prioritization, Bioinformatics 23 (2007) i125–32.

6. Tables

Table 1: Basic statistical information about our annotated dataset.

Number of Proteins	10,282
Min Degree	1
Max Degree	272
Average Degree	9.39
Number of Proteins with no Function	1519
Average Number of Functions for each Protein	10.40

Table 2: 20 drugs with the highest number of seed proteins.

DrugName	Seed Count
NADH	77
L Glutamic Acid	49
Adenosine triphosphate	30
Alpha D Mannose	29
Adenosine 5 Diphosphate	25
Clozapine	24
Quetiapine	24
Aripiprazole	23
Ziprasidone	23
Olanzapine	23
Nicotinamide Adenine Dinucleotide	23
Pyridoxal Phosphate	23
Amitriptyline	19
Marimastat	19
Glycine	18
Doxepin	18
Bromocriptine	18
Temazepam	17
Pergolide	17
Lorazepam	17

Table 3: 13 clusters extracted from the pruned network using MCL clustering [64].

Cluster ID	Cluster Members
Cluster 1	Nitrazepam, Halazepam, Oxazepam, Heptabarbital, Hexobarbital, Ethchlorvynol, Nicotine, Dopamine, Barbital, Talbutal, Barbituric-acid-derivative, Aprobarbital, Adinazolam, Butalbital, Glycine, Galantamine, Lorazepam, Quazepam, Clotiazepam, Flurazepam, L-Glutamic-Acid, Temazepam, Methylphenobarbital, Primidone, Butethal, Butabarbital, Pentobarbital, Amobarbital, Midazolam, Halothane, Amoxapine, Vitamin-E, Metharbital, Secobarbital, Phenobarbital, Clonazepam, Clobazam, Fludiazepam, Chlordiazepoxide, Bromazepam, Thiopental, Eletriptan, Estazolam, Triazolam, Alprazolam, Prazepam, Succinic-acid, Diazepam, Cinolazepam, Clorazepate
Cluster 2	Ropinirole, Olanzapine, Promazine, Lisuride, Pramipexole, Disopyramide, Minaprine, Clozapine, Droxidopa, Carvedilol, Apomorphine, Epinephrine, Promethazine, Yohimbine, Nortriptyline, Sertindole, Aripiprazole, Pergolide, Quetiapine, Norepinephrine, Bromocriptine, Acepromazine, Chlorpromazine, Trimipramine, Chlorprothixene, Cabergoline, Methotrimeprazine, Ziprasidone, Propiomazine, Thioridazine, Amitriptyline, Doxepin, Thioproperazine, Paliperidone, Maprotiline, Desipramine, Risperidone, Ergotamine, Imipramine
Cluster 3	Trastuzumab, Basiliximab, Adalimumab, Tositumomab, Alemtuzumab, Muromonab, Alefacept, Natalizumab, Rituximab, Etanercept, Gemtuzumab-ozogamicin, Ibritumomab, Bevacizumab, Cetuximab, Palivizumab, Daclizumab, Abciximab, Efalizumab
Cluster 4	Alpha-D-Mannose, Coagulation-Factor-IX, Vitamin-C, Urokinase, Marimastat, Antihemophilic-Factor, Menadione, Benzamidine, Drotrecogin-alfa, L-Carnitine, Tenecteplase
Cluster 5	Heme, Famoxadone, 5-n-undecyl-6-hydroxy-4-7-dioxobenzothiazole, Ubiquinone-2, Minocycline, 2-Nonyl-4-hydroxyquinoline-N-oxide, NADH
Cluster 6	Staurosporine, Sorafenib, Imatinib, Dasatinib, Palifermin, Sunitinib, Adenosine-triphosphate
Cluster 7	Icosapent, Acitretin, Alitretinoin, Sulfasalazine, Genistein
Cluster 8	Paroxetine, Tramadol, Pseudoephedrine, Ephedra, Methamphetamine
Cluster 9	Nefazodone, Clomipramine, Cocaine, 3-4-Methylenedioxymethamphetamine
Cluster 10	Puromycin, Cladribine, Alpha-Hydroxy-Beta-Phenyl-Propionic-Acid, Anisomycin
Cluster 11	Amlodipine, Nimodipine, Mibefradil
Cluster 12	Verapamil, Nifedipine
Cluster 13	Hydroflumethiazide, Zonisamide

Table 4: 10 highest-ranked drug target proteins predicted for each cluster shown in table 3.

Cluster ID	Predicted drug target proteins
Cluster 1	HTR3A,CHRNE,CHRNG,CHRND,FXYD1,CRELD2,GPR26,HAAO, GRID2IP,CLIC6
cluster 2	ADORA1,ADORA2A,GRIN2A,GRIN2B,EDNRA,CHRNA4,CHRNA3, CACNA1C,P2RX4,TACR1
cluster 3	DPP4,CCL5,TFRC,CD86,PVRL1,F2,VCAM1,ADIPOQ,CR2,IL1B
cluster 4	APOE,THBS1,ALS2,FGB,ANXA1,C4BPB,TGFB1,SERPINA10,MGST3,GAS
cluster 5	CYP1A1,FH,PDK4,CYP2C9,CYP2C19,CYP1A2,CYB5R1,PC,CYP2E1,FDX1
cluster 6	FRAP1,ACVR2A,ACVR2B,ENG,AKT2,ACVR1C,ABCB4,TGFBR2, CFTR,BMPR2
cluster 7	CTNNB1,BAAT,IRX4,ALOX5AP,REXO4,XPMC2H,TRIM59,TRPV2, SLC30A9,ZNF398
cluster 8	SLC6A1,DRD4,DRD2,DRD1,DRD3,BCL2,OXTR,DNAJB4,NDUFB10,ABR
cluster 9	HTR1B,SLC6A1,DRD2,DRD3,GRIN2A,OXTR,DRD1,BCL2,PSEN1,CHRNB2
cluster 10	RPS11,RPS7,RPL35,RPL32,MRPS5,RPL36A,RPL30,RPS13,RPL34,RPL29
cluster 11	CACNA1A,RYR1,KCNMA1,CACNG2,CACNA1E,SLC8A1,CACNA2D4, CATSPER1, KCNJ12,SCN8A
cluster 12	KCNMA1,RYR1,CACNA1E,CACNA2D4,KCNA2,KCNH1,ALG10B,KCR1, KCNIP2,KCNQ1
cluster 13	KCNJ12,CACNA1A,SLC8A1,SCN10A,CACNA1S,SCN8A,CA9,CACNA1C, KCNQ1,CACNB2

7. Figures

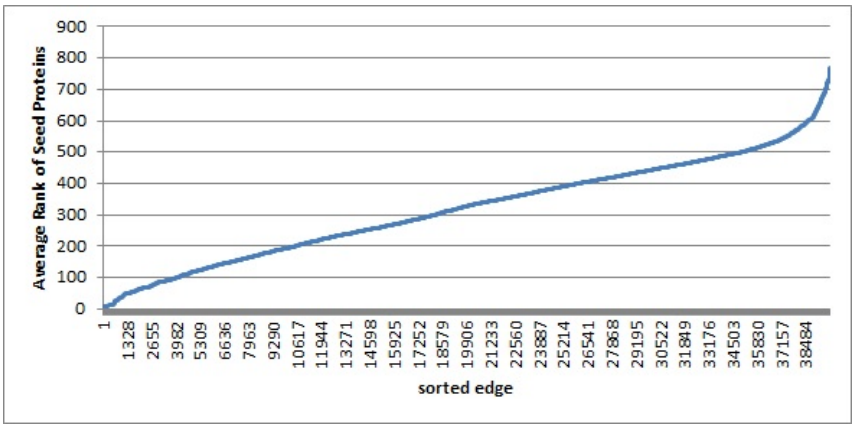

Figure 1: HDN Edge Distribution. The X axis shows the 39800 edges in the HDN and the Y axis shows, for each edge $d_i \rightarrow d_j$, the average rank of the seed proteins of d_j (the smaller, the better). There are two turning points in the curve, roughly at 3% and 86% of all edges.

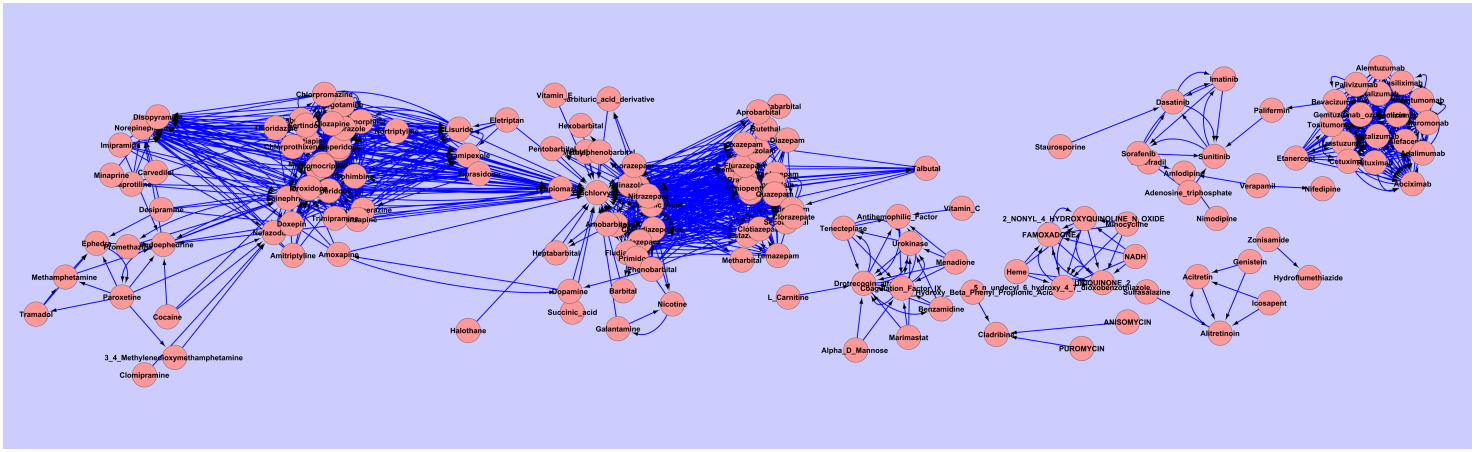

Figure 2: Pruned Human Drug Network including only 1328 highest relationship (3% of the original network). The biological significance of the observed findings is discussed in Section 4.2.

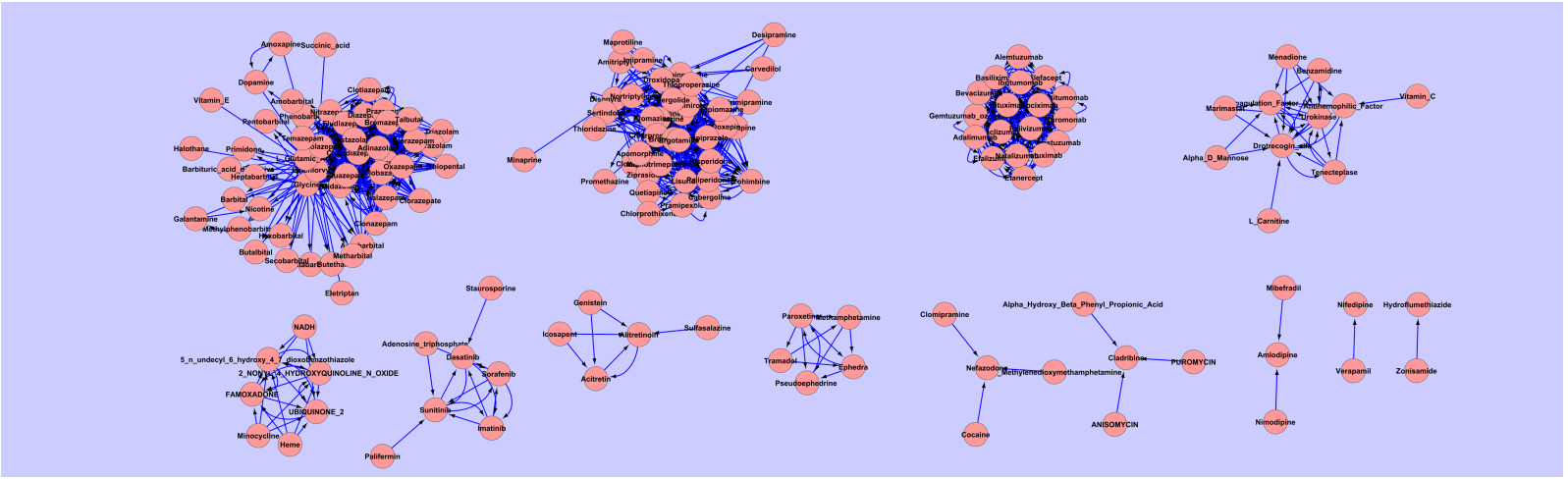

Figure 3: Clustering the pruned HDN using MCL Algorithm [64]. Section 4.2 discusses the clusters in details.

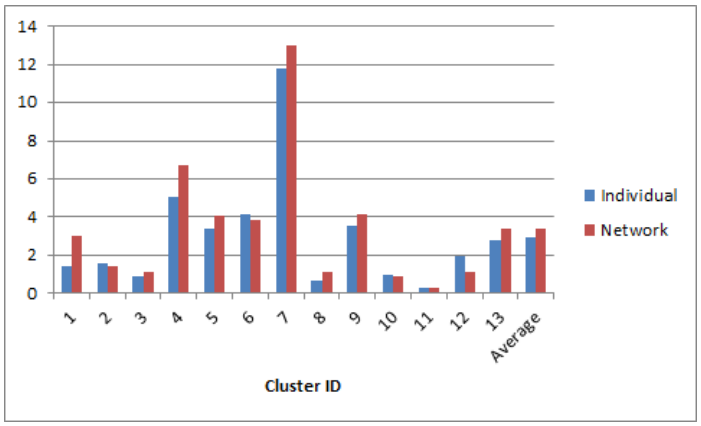

Figure 4: Comparing Individual approach with Network approach with respect to average rank of seed proteins (The smaller, the better). The Network approach outperforms the Individual approach in clusters 2, 6, 10 and 12 with 39, 7, 4 and 2 drug members.